Grid Technology as a Cyberinfrastructure for Earth Science Applications

Thomas H. Hinke
NASA Advanced Supercomputing Division
NASA Ames Research Center
Moffett Field, CA 94035

Abstract - This paper describes how grids and grid service technologies can be used to develop an infrastructure for the Earth Science community. This cyberinfrastructure would be populated with a hierarchy of services, including discipline specific services such those needed by the Earth Science community as well as a set of core services that are needed by most applications. This core would include data-oriented services used for accessing and moving data as well as compute-oriented services used to broker access to resources and control the execution of tasks on the grid. The availability of such an Earth Science cyberinfrastructure would ease the development of Earth Science applications. With such a cyberinfrastructure, application workflows could be created to extract data from one or more of the Earth Science archives and then process it by passing it through various persistent services that are part of the persistent cyberinfrastructure, such as services to perform subsetting, reformatting, data mining and map projections.

I. Introduction

The Earth Science community uses data that is stored in widely distributed Earth Science archives provided by the various Distributed Active Archive Centers (DAACS) supported by NASA and other agencies, as well as the archives supported by the Earth Science Information Partners (ESIPS). Many scientists and others who use Earth Science data have a need for various tools or services to transform this data into a form that they can readily use. Examples of such tools or services are subsetting tools, reformatting tools, map projection services and data mining services. Today individual groups build their own tools or procure and install tools that have been developed by others.

Since the data, the applicable tools and the expertise to maintain and use these tools may be widely distributed across geographic and administrative boundaries, it would be desirable to have an Earth Science cyberinfrastructure to provide a unifying structure for users to access and process data in a seamless manner, irrespective of the location of the data or the tools. This infrastructure could also facilitate collaboration between distributed groups. In support of this vision, this paper describes how grids and grid service technology can provide an appropriate cyberinfrastructure to integrate data sources, computational resources and services into a system that can easily support the needs of the Earth Science community.

The next section will provide an overview of grids, grid service technologies and the work that NASA has done that is relevant to creating an Earth Science cyberinfrastructure. The following section will then look at some of the applications of this technology that has supported or could support Earth Science applications. The final section will describe a possible roadmap for developing an Earth Science cyberinfrastructure, which could permit the Earth Science community and the various consumers of Earth Science information and expertise to more easily and quickly realize their objectives.

II. Grids And Grid Services

The section will describe the features of grids and grid service technology that are relevant to the development of an Earth Science cyberinfrastructure and how the current work in grid services has it roots in both grid technology and web service technology. It will also discuss the work that NASA has done that can contribute to this work.

A. Grid Technology

Grids are a key element of a cyberinfrastructure since they consist of middleware tools that provide secure and seamless access to distributed computational resources, data archives and scientific instruments or sensors. Grids provide tools to remotely execute jobs and applications on widely distributed computers, to transfer data among widely distributed data storage resources and to do all of this securely.

A key aspect of grid security is that it creates a single sign-on environment. This means that a user only needs to authenticate himself to the first grid resource that he uses. The grid will then allow him to access and use any other grid resource for which he is authorized, without having to do any additional authentication.

In systems such as the Globus grid toolkit [1] (the *defacto* grid middleware standard), a user's identify is represented by an X-509 certificate that is digitally signed by a trusted certification authority. When a user first authenticates himself to the grid by providing a secret pass phrase (like a password), a secure, limited lifetime proxy certificate is created from the certificate, to be used to authenticate the user to every grid resource that he uses during the session.

Each grid resource will authenticate the user based on the identity contained in the proxy certificate and will map the user's grid identity (as indicated by the grid proxy) into the local identity by which the user is known on that resource.

A grid is formed by a layer of middleware that binds all of the distributed resources of a grid into a system that can support transparent, secure access to all of these resources. Ian Foster, one of the originators of Globus grid technology has suggested a three-point checklist [2] for determining if something is a grid. According to his definition, "… a grid is a system that:

- Coordinates resources that are not subject to centralized control ….
- Using standard, open, general-purpose protocols and interfaces ….
- To deliver nontrivial qualities of service …."

An important feature of grids is that they provide the ability to remotely execute applications on the computational resources to which a user has legitimate access, with the grid middleware handling the interface to the remote system's job launching mechanisms. Another capability supported by grids is the ability to securely transfer files, including what are called third-party transfers in which the user can request the transfer between any two arbitrary grid resources, irrespective of where the user is currently logged in.

A number of major grids exist in the world today and the number is growing. Some of the more significant include the following:

- Information Power Grid developed by NASA to be a testbed for NASA oriented grid developments [3]
- TeraGrid funded by the National Science Foundation to support open scientific research [4]
- DOE Science Grid, whose goal is to provide a cyber infrastructure that spans Department of Energy resources [5]
- NEESgrid that provides access to computing resources and research equipment for earthquake research in the US [6]
- UK e-Science grid, that supports science in the United Kingdom [7]
- European Data Grid under the European Union [8]

In addition, grid projects exist in the Scandanavian countries (NorduGrid), the Asian-Pacific basin (Asian Pacific Grid), Korea, Japan and China.

One of the forces facilitating the growth of grid technologies is the development of grid standards through international standards organizations such as the Global Grid Forum (GGF) [9]. This organization is patterned after the Internet Engineering Task Force (IETF) [10], which develops many of the network standards that facilitate the growth of network technology. GGF's charter is to develop best practices and standards for grid technologies. Its objective is the development of open standards that can be used by all interested vendors to develop grid middleware. Much of the GGF standards' work is centered around the Open Grid Service Architecture (OGSA), which envisions a grid fabric composed of various core grid services that are required for constructing most grid system, and which can provide the building blocks that will ease the construction of application-oriented services. [11] While the standard is still in draft form, some of the types of functional capabilities that these core services will address are resource discovery and brokering, accounting, data sharing and monitoring.[12]

The nature of the service infrastructure that will comprise OGSA and the services that will be built on top of it will be described in the next section, which provides an overview of the grid service approach.

B. Grid Service Technology

The concept of a web service [13] was developed by the commercial sector as a means to provide the ability of software to access data, in the form of a service. While the well-known web page technology has provided a means for humans to easily access information on web pages using browsers, it was not designed to make it easy for software to conveniently access data from remote sites. To answer this need, web services were developed as a means for software to be able to easily access remote data through the use of an encapsulating service.

Seeing the value of this technology, the grid community began to develop a similar technology in the form of an Open Grid Service Infrastructure (OGSI) [14]. In late 2003, a group comprised of a number of the leading commercial vendors and grid developers recommended a standard called the Web Services Resource Framework (WS-RF)[15], which unified grid services and web services. This standard has been submitted to OASIS (Organization for the Advancement of Structured Information Standards) [16], which is standardizing commercial web services. GGF will then build its OGSA standards on top of this grid services fabric that is provided by WS-RF.

As an initial reference implementation of grid technologies using the proposed WS-RF standards, the Globus project will be releasing a version of their Globus grid tool kit in the form of Globus Toolkit 4 in 2004.

The now combined grid and web service technology, which will be referred to simply as grid services in the

remainder of this paper, could revolutionize the ability of developers to easily construct extensible data handling and processing systems, by providing seamless access to various types of resources including computational resources, data archives, scientific sensors and the services to process data into some desired product, which itself could be formulated as a service.

One of the key capabilities of grid services is the ability to describe service interfaces. Using a Web Service Description Language (WSDL), the developer of a grid service can describe the interface to that service including the manner in which the interface has to be called to utilize the desired service. This means that a number of different groups could each provide services that perform the same general type of function, but they would not be constrained to all use the same interface.

Applications or other services that were looking for a desired service would contact what is called a UDDI (Universal Description, Discovery and Integration) registry to find out what services existed to provide some needed capability and the nature of the WSDL described interface, which indicates how that service is to be accessed. At this point in the development of grid service technology, the user's application would have to already know how to interact with each type of service. Ultimately there should be shared ontologies covering each of the domains addressed by the so that clients can adapt to use a variety of similar services.

C. NASA's Contributions to Grids and Grid Services

The NASA Ames Research center has been working on grid technologies for a number of years, with the objective of developing grid technologies that will increase the intelligence of the grid and provide re-usable components in support of the many grids that may be used throughout NASA in the future; including perhaps some future Earth Science Grid. The following services have been developed to increase the intelligence of grid processing:

- A resource broker service that select the "best" set of resources based on user requirements, [17]
- A naturalization service that automatically tailors the processing environment on grid resources so that programs such as data mining systems that are staged to them will run with the necessary shared libraries and environment settings. [18]
- An execution service that can autonomously manage the user's job as it moves through stages execution, and

Work is currently underway to develop the following capabilities:

- A dynamic access service [19] that will permit users to instantly, but with proper accountability, access needed computational resources across administrative boundaries, without having to have pre-established accounts on these machines, using policy-based authorization, and
- A grid-based data handling system to build on research that has been done in the wider grid community to make grid-accessible metadata catalogs available to the NASA-oriented grid processing.

Work is also underway focused on increasing the intelligence of grid management and assessment services. This includes the development and adaptation of tools and processes that enable rapid integration of grid technologies for NASA applications.

Figure 1 shows a grid service architecture that includes the major areas where work is being done to increase the intelligence and capabilities of the grid middleware. As noted, some of this work is being done in-house, some is being partially supported by NASA and other work is available from the wider grid community, which tends to develop open source software that is available to all.

III. Earth Science Use Of Grids And Grid Services

Grids and grid services are beginning to be used for Earth Science applications. In addition, their use for other applications may also have applicability to the Earth Science community. This section will discuss a number of Earth Science and Earth Science related projects that illustrate various facets of this technology.

A. Data Mining and Multi-center Data Production

In this first application, the general remote processing and data movement capabilities of the grid are used, with grid services providing part of the core grid infrastructure that facilitated the construction of the system. In this application, the grid, in collaboration with a grid-enabled data mining system called the Grid Miner [20] (based on a stand-alone data mining system funded by a NASA NRA and created by the University of Alabama in Huntsville) has been used to create a system that uses the grid to produce Earth Science data products that involve data from multiple archives. [21] This system consults a Resource Broker grid service to determine where the Grid Miner should run, based on supplied criteria, such as the required operating system. The system then selects one of the systems and uses an Execution grid service to stage a thin Grid Miner agent and an associated mining plan to the site selected to perform the mining. At this site, the Grid Miner agent consults the mining plan to determine what mining

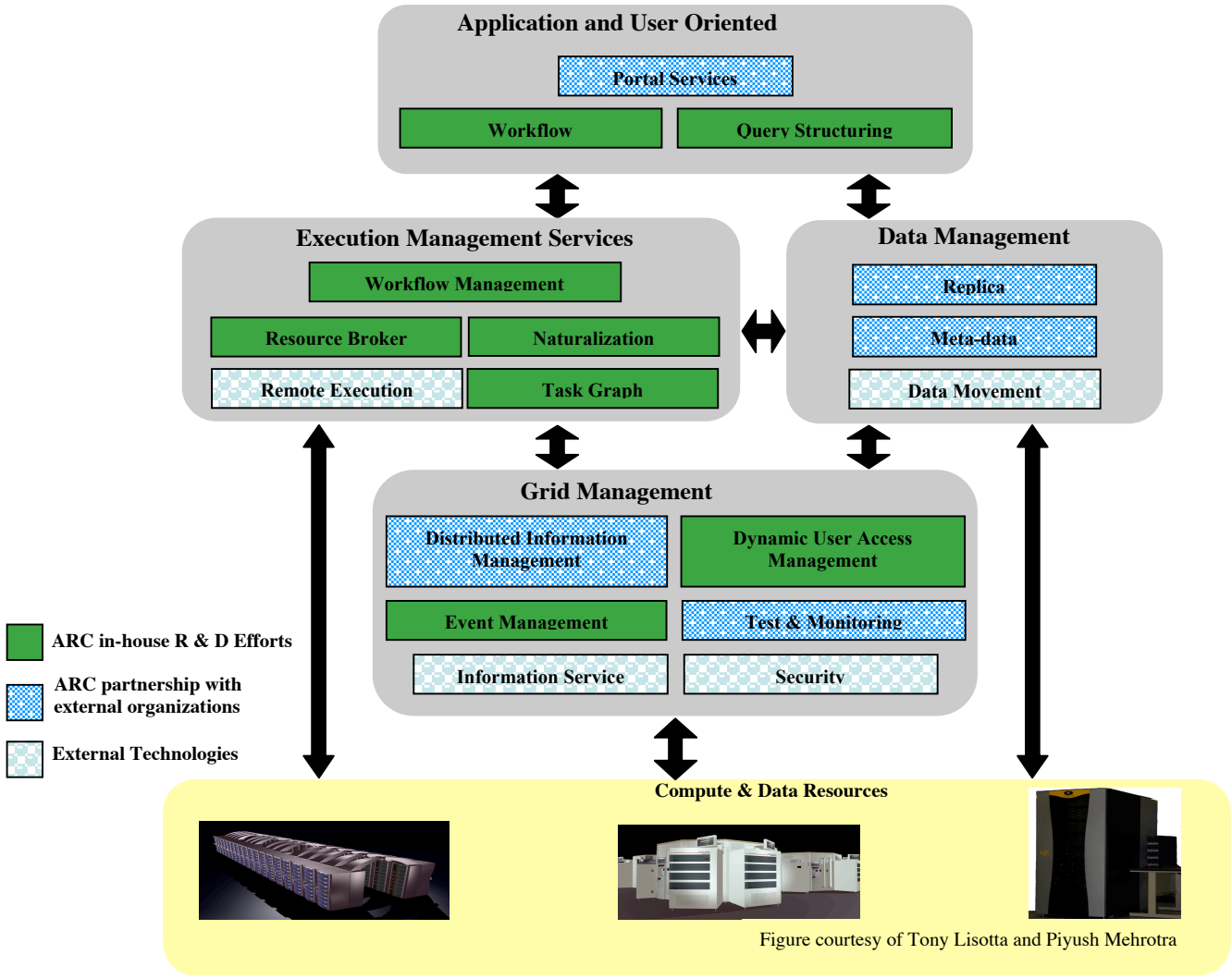

Fig 1. The development of NASA grid components is a collaboration of NASA, NASA partners and external, open source development projects.

operations are to be performed and then uses the grid data transfer capability to transfer the necessary mining code from a grid-accessible repository to the mining site. It then consults a mining database to determine the remote location of the files to be mined, and uses the data transfer capabilities of the grid to transfer these to the mining site.

Figure 2 shows a simplified description of the architecture. In this architecture, TMI (TRMM [Tropical Rainfall Measuring Mission] Microwave Imager data, which was originally acquired from the NASA Goddard Space Flight Center's Distributed Active Archive Center, was extracted from a grid-enabled tertiary storage at NASA Ames Research Center. The grid-enabled technology that was used to support this seamless access from tertiary storage to mining site was the Storage Resource Broker, which was developed by the San Diego Supercomputing Center. [22]

In this application, the Grid Miner mines TMI data for mesoscale convective systems (MCS), which is a severe storm, using an algorithm originally suggested by [23]. The results of the mining are XML (extensible markup language) encoded polygons that circumscribe an MCS. These XML polygons are transferred using the grid's data transfer capability to the Atmospheric Sciences Data Center at NASA Langley Research Center. There they are used to subset CERES (Clouds and the Earth's Radiant Energy System) data. The result of this application is that one now has CERES subsets that correspond to a mesoscale convective system.

It should be noted that in this case, even though the data sets are distributed in different archives, they were both captured from instruments that were on the same satellite, so the data that is involved in this data production process

were all captured at the same time. Additional details about this work are described in [9].

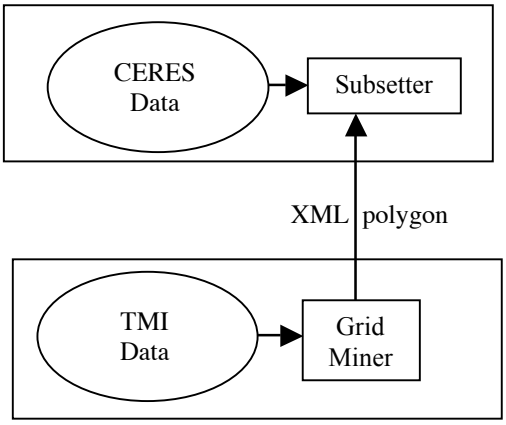

Fig. 2 - Data mining and the grid are used to produce a data product that involves data from multiple data center

An example of the grid providing parallel mining was shown by a later application that took the 14 orbits that constituted one day of TMI data and spread them across a number of different grid processors that had been identified by the broker, with each processor running its own instance of the Grid Miner. The mining results from each processor were then sent to another node running yet another instance of the grid miner that had a mining plan that fused the different orbit results into a unified result, which was then used to drive a visualization.

B. Earth Science Grid Service Example

This next application illustrates the use of a grid to host an Earth Science oriented service, called the NASA Web GIS Software Suite (NWGISS), which is an example of the type of services that an Earth Science cyberinfrastructure would be expected to support. This service, developed at George Mason University under the direction of Dr. Liping Di, seeks to build on the work of the Open GIS (Geographical Information System) Consortium (OGC) to serve up Earth Science data that complies with the OGC standards. In this form, the data can be ingested into commercial GIS systems, whereas the original Earth Science data that is stored in some format, such as HDF-EOS (which is extensively used by EOSDIS -- Earth Observing System Data and Information System) cannot be ingested into a GIS system. [24]

In addition to the ability to supply OGC-compliant Earth Science data, this system also provides a grid-accessible catalog that can be used for discovering data of interest. Once the user has used the catalog to find the data that he needs, he can then order it from NWGISS system and receive it over the grid.

NWGISS can be used to access data that is stored on a system that contains the NWGISS processing software as well as a system that contains the data but does not contain any NWGISS processing software. In this later case, NWGISS uses the gridFTP capability to transport the data to a processor that contains the NWGISS processing service, where it is processed and sent to the user. In this case, the NWGISS system itself is comprised of services that are distributed and uses the grid to gain access to remote services that are required for processing the data.

The NWGISS service itself could be a component of a more complex system. This is the value of a grid-service based cyberinfrastructure, in that services can form the components for more complex services, which can then form the components for even more complex services. The service-based cyberinfrastructure is the enabling technology that permits each of these services to be invoked in support of other services.

C. International Use of Grid Technologies

To provide an idea of how grid technologies are being used for Earth Science internationally, this section will take a brief look at the grid testbed being assembled by a working group of the Committee on Earth Observing Satellites (CEOS). CEOS is an international organization of representatives from countries that use satellite data. It is developing a CEOS grid testbed under the direction of the CEOS Working Group on Information Systems and Services (WGISS) that represents organizations that run Earth Science data centers. This effort began with an initial workshop in 2002 to explore how grid technologies could be of value to the CEOS members. Based on the results from the workshop, work on the CEOS grid testbed began in late 2002 and continues to this day to evaluate the applicability of grid technologies within various Earth Science applications of interest to the WGISS members. The following provides a brief summary of the of the organizations and projects that comprise the CEOS grid testbed:

- US Geological Survey: Explore use of GRID technologies for the delivery/reception of Earth Science data,
- National Oceanic and Atmospheric Administration: Provide access to climate and numerical weather prediction models for analysis and intercomparison, foster research to study complex earth systems using collections of distributed data,
- European Space Agency: Generic infrastructure to allow seamless plug-in of specific data handling and

application services to support on-demand user-driven data integration,

- NASA Goddard Space Flight Center and George Mason University: Integration of Grid and Open GIS Consortium (OGC) Web Services (previously discussed in this paper)
- George Mason University: Provide the ability to advertise and deliver virtual datasets
- University of Alabama in Huntsville: Compute-intensive data mining and machine learning applications in the Earth Sciences
- DutchSpace and European Space Agency/European Space Research Institute: Simulators of Earth observing instruments and data processing software working together using Computational Grid technologies

In support of the CEOS grid testbed, NASA Ames is providing grid consulting to the group as a whole and has supplied CEOS grid participants with both X509 user certificates, by which the identity of user's can be authenticated, and X509 host certificates, by which the identity of the host computers that comprise the CEOS grid testbed can be authenticated. Since establishing a certificate authority to issue user and host certificates can be a fairly time-consuming undertaking, the use of certificates issued by an organization that has already established a certificate authority can save time and money.

It should be noted, that the issuing of a certificate does not imply any access to any resource. Certificates are designed to support authentication. Authorization, which determines who can access what resource, is a step that relies on authentication to determine the identify of the user that is attempting to access the resource, but then goes beyond authentication to determine whether or not a user is permitted to access a particular resource based on some properties of the resource and user.

D. Space-based Data Distribution Service

As a final application area, this section will look at a grid service that distributes real time data flow from space to provide yet another building block as part of a cyberinfrastructure. Under the direction of Robert Bradford at NASA's Marshall Space Flight Center, a Space Development and Operations Grid (SpaceDOG) is being developed. One of the capabilities being developed for SpaceDOG in collaboration with personnel from NASA Ames is the creation of a telemetry grid-service that will distribute International Space Station (ISS) telemetry data as well as telemetry data associated with ISS payloads.

With such a telemetry grid service, principal investigators will be able to use this service to feed data directly to their applications or to applications that are being developed by others that use the data from the particular PI's experiment. In addition, this telemetry grid-service will be able to distribute general data about ISS, which could be used by the PIs or others interested in ISS operations. The point is that this now becomes a general service with a well-defined WSDL interface that can be used to support many other services. As the workload increases, multiple instances of this service could be fielded.

As currently configured the grid does not extend to the ISS. The telemetry grid service has been constructed by wrapping a legacy application with grid service code, so that it projects a grid service interface to the grid and responds to those applications that know how to interface to a grid service. However, there is no inherent reason why grids could not be extended into space, such that the various sensors on space-based platforms could become a node on an Earth-Space grid,

With such an Earth-Space grid, each of the sensors on Earth-oriented satellite could be a node on the grid, with their data transport being handled by the grid. In addition, the potential exists for this Earth-Space grid to be used for controlling the sensors. In previous collaborative work between NASA Ames, the San Diego Supercomputer Center and the University of California at San Diego, a remote tele-science capability was supported. In this application, a NASA scientist at Wallops Island, Virginia used a grid-enabled portal (developed by the San Diego Supercomputer Center) to control an electron microscope at the University of California at San Diego, with the data from that work being shipped over the grid to a storage system at Ames.

This same technology could be applied to a Earth-Space grid to make real-time data from a satellite-based service an integral part of some workflow that consisted of both space-based and terrestrial grid services. Issues to be addressed are the longer latency than are found on terrestrial grids, the potentially lower data rates and the periodic dropping of connectivity as the satellite leaves the areas where the data can be directly down-linked to the terrestrial portion of the grid.

IV. Earth Science Cyberinfrastructure Roadmap

A possible roadmap to an Earth Science cyberinfrastructure could include the following steps:

Phase 1: Provide grid access to the Earth Science archives e.g., EOSDIS Distributed Active Archive Centers and Earth Science Information partners (ESIPs)

- Step 1.1: Provide grid access to disk resident data (data pools) associated with each archive.

- Step 1.2: Provide grid access to data resident on mass storage systems. While this capability exists with current technologies (such as the grid-enabled Storage Resource Broker), approaches will have to be developed for moderating this type of access since the existing data paths out of mass storage systems are narrow, and these paths could be overwhelmed if too many users tried to move data at the same time.

- Step 1.3: Provide grid accessible metadata catalogs to permit applications to automatically locate desired data based on data criteria and then update the catalog with metadata for newly created products.

Phase 2: Provide grid services that can perform core Earth Science processing

- Step 2.1: Provide persistent services to perform common Earth Science functions such as subsetting, reformatting, transforming data into desired map projections, or data mining

- Step 2.2: Integrate archives and services into a seamless infrastructure that permits scientists to develop workflows to extract desired data from the archive, move the data through the various processing services that are needed to produce the desired results, store the new product back on the grid-accessible archive, and then update the catalog with metadata associated with the newly created data.

Phase 3: Extend the Earth Science cyberinfrastructure into space, making the various satellite-based sensors part of the cyberinfrastructure.

- Step 3.1: Extend the cyberinfrastructure fabric into space, which under the current state-of-the art for grids will require that the Internet Protocols (e.g., IP) be extended into space and onto satellites and issues of reduced data rates, periodic disconnects and high latency be addressed.

- Step 3.2: Implement sensors as grid services

In conclusion, grid technologies are currently available that can provide the basis for implementing an Earth Science cyberinfrastructure. Initial version of this can be used by early adopters, but as the system is refined through use, it can be gradually opened to a wider group of Earth Science users and applications.

ACKNOWLEDGEMENT

I would like to acknowledge the Judith Utley from NASA Ames, who is actively involved with the CEOS grid testbed project and provided the information on which the CEOS portion of this paper is based. I would also like to acknowledge valuable suggestions made by Dr. Piyush Mehrotra and Dr. Warren Smith on earlier versions of this paper.

REFERENCES

[1] I. Foster, C. Kesselman. *Intl J. Supercomputer Applications* , 11(2):115-128, 1997.

[2] Ian Foster, "What is the Grid? Three Point Checklist," Grid Today, Vol. 1 No. 6, July 22, 2002.

[3] W. E. Johnston, D. Gannon, B. Nitzberg, "Grids as Production Computing Environments: The Engineering Aspects of NASA's Information Power Grid ,"Eighth IEEE International Symposium on High Performance Distributed Computing, Redondo Beach, CA, Aug. 1999

[4] D.A. Reed, Grids, "The TeraGrid and beyond," Computer , Volume: 36 , Issue: 1 , Jan. 2003

[5] The DOE Science Grid: FY2003 Accomplishments, http://doesciencegrid.org/Grid/papers/DOEScienceGrid-SciDAC.2.pager.pdf, 2003

[6] NEESgrid, http://www.neesgrid.org

[7] A.J.G. Hey and A.E. Trefethen, "The UK e-Science Core Program and the Grid," Future Generation Computer Systems, vol. 18, no. 8, Oct. 2002, pp. 1017–1031.

[8] European Data Grid, http://web.datagrid.cnr.it

[9] Global Grid Forum, www.ggf.org

[10] Internet Engineering Task Force, http://www.ietf.org/

[11] I. Foster, C. Kesselman, J. Nick, S. Tuecke , "The Physiology of the Grid: An Open Grid Services Architecture for Distributed Systems Integration." Open Grid Service Infrastructure WG, Global Grid Forum, June 22, 2002. (http://www.globus.org/research/papers.html)

[12] I. Foster, D.Gannon, H. Kishimoto, "Open Grid Service Architecture," GGF draft ggf-ogsa-spec-01, Global Grid Forum, March 2004.

[13] W3C Web Service Activity, http://www.w3.org/2002/ws/

[14] GGF OGSI Working Group, https://forge.gridforum.org/projects/ogsi-wg 1

[15] Karl Czajkowski, Donald F Ferguson, Ian Foster , Jeffrey Frey , Steve Graham , Igor Sedukhin, David Snelling , Steve Tuecke and William Vambenepe, The

WS-Resource Framework, Version 1.0, March 5, 2004. IBM Web Site, http://www-106.ibm.com/developerworks/library/ws-resource/ws-wsrf.pdf.

[16] OASIS Web Service Resource Framework Technical Committee, http://www.oasis-open.org/committees/tc_home.php?wg_abbrev=wsrf.

[17] Paul Kolano, "Surfer: An Extensible Pull-Based Framework for Resource Selection and Ranking," 4th IEEE/ACM Intl. Symposium on Cluster Computing and the Grid, Chicago, April 2004.

[18] Paul Kolano, "Facilitating the Portability of User Applications in Grid Environments," 4th International Conference on Distributed Applications and Interoperable Systems, Paris, France, November 2003

[19] Rebekah Lepro, "Cardea: Providing Support for Dynamic Resource Access in a Distributed Computing Environment," 19th Annual Computer Security Applications Conference (as work in progress), Las Vegas, Nevada, Dec. 2003

[20] Thmas H. Hinkeand Jason Novotny, "Data Mining on NASA's Information Power Grid," Proceedings Ninth IEEE International Symposium on High Performance Distributed Computing , Pittsburgh, Pennsylvania, August, 2000.

[21] Bruce R. Barkstrom, Bruce R., Thomas H. Hinke, Shradha Gavali, Warren Smith, William J. Seufzer, Chaumin Hu, David E. Cordner, "Distributed Generation of NASA Earth Science Data Products," Journal of Grid Computing , 1(2): 101-116, 2003

[22] C. Baru, R. Moore, A. Rajasekar, and M. Wan, "The SDSC Storage Resource Broker," Proceedings of the CASCON'98, Toronto, Canada, 1998.

[23] K. I. Devlin, Application of the 85 GHz Ice Scattering Signature to a Global Study of Mesoscale Convective Systems. Master's thesis, Meteorology, Texas A&M University, August 1995.

[24] L. Di, A. Chen, W. Yang, P. Zhao, "The Integration of Grid Technology with OGC Web Services," Global Grid Forum, Workshop on Grid Applications and Programming Tools, Seattle,June 2003